A Large-Scale Evaluation of Acoustic and Subjective Music-Similarity Measures

Adam Berenzweig,* Beth Logan,† Daniel P.W. Ellis,* and Brian Whitman†‡
*LabROSA
Columbia University
New York, New York 10027 USA
alb63@columbia.edu
dpwe@ee.columbia.edu
†HP Labs
One Cambridge Center
Cambridge, Massachusetts 02142–1612 USA
beth.logan@hp.com
‡Music, Mind & Machine Group
MIT Media Lab
Cambridge, Massachusetts 02139–4307 USA
bwhitman@media.mit.edu

A valuable goal in the field of Music Information Retrieval (MIR) is to devise an automatic measure of the similarity between two musical recordings based only on an analysis of their audio content. Such a tool—a quantitative measure of similarity—can be used to build classification, retrieval, browsing, and recommendation systems. To develop such a measure, however, presupposes some ground truth, a single underlying similarity that constitutes the desired output of the measure. Music similarity is an elusive concept—wholly subjective, multifaceted, and a moving target—but one that must be pursued in support of applications to provide automatic organization of large music collections.

In this article, we explore music-similarity measures in several ways, motivated by different types of questions. We are first motivated by the desire to improve automatic, acoustic-based similarity measures. Researchers from several groups have recently tried many variations of a few basic ideas, but it remains unclear which are best-suited for a given application. Few authors perform comparisons across multiple techniques, and it is impossible to compare results from different authors, because they do not share the required common ground: a common database and a common evaluation method.

Of course, to improve any measure, we need an evaluation methodology, a scientific way of determining whether one variant is better than another. Otherwise, we are left to intuition, and nothing is gained. In our previous work (Ellis et al. 2002), we have examined several sources of human opinion about music similarity, with the impetus that human opinion must be the final arbiter of music similarity, because it is a subjective concept. However, as expected, there are as many opinions about music similarity as there are people to be asked, and so the second question is how to unify the various sources of opinion into a single ground truth. As we shall see, it turns out that perhaps this is the wrong way to look at things, and so we develop the concept of a "consensus truth" rather than a single ground truth.

Finally, armed with these evaluation techniques, we provide an example of a cross-site evaluation of several acoustic- and subjective-based similarity measures. We address several main research questions. Regarding the acoustic measures, which feature spaces and which modeling and comparison methods are best? Regarding the subjective measures, which provides the best single ground truth, that is, which agrees best on average with the other sources?

In the process of answering these questions, we address some of the logistical difficulties peculiar to our field, such as the legal obstacles to sharing music between research sites. We believe this is one of the first and largest cross-site evaluations in MIR. Our work was conducted in three independent labs (LabROSA at Columbia, MIT, and HP Labs in Cambridge), yet by carefully specifying our evaluation metrics, and by sharing data in the form of derived features (which presents little threat to copyright holders), we were able to make fine dis-

Computer Music Journal, 28:2, pp. 63–76, Summer 2004

tinctions between algorithms running at each site. We see this as a powerful paradigm that we would like to encourage other researchers to use.

Finally, a note about the terminology used in this article. To date, we have worked primarily with popular music, and our vocabulary is thus slanted. Unless noted otherwise, when we refer to "artists" or "musicians" we are referring to the performer, not the composer (which frequently are the same anyway). Also, when we refer to a "song," we mean a single recording of a performance of a piece of music, not an abstract composition, and also not necessarily vocal music.

This article is organized as follows. First we examine the concept of music similarity and review prior work. We then describe the various algorithms and data sources used in this article. Next, we describe our evaluation methodologies in detail and discuss issues with performing a multi-site evaluation. Then we discuss our experiments and results. Finally, we present conclusions and suggestions for future directions.

Music Similarity

The concept of similarity has been studied many times in fields including psychology, information retrieval, and epistemology. Perhaps the most famous similarity researcher is Amos Tversky, a cognitive psychologist who formalized and studied similarity, perception, and categorization. Tversky was quick to note that human judgments of similarity do not satisfy the definition of a Euclidean metric, as discussed below (Tversky 1977). He also studied the context-dependent nature of similarity and noted the interplay between similarity and categorization. Other notable work includes Goldstone, Medin, and Gentner (1991) and the music psychology literature—e.g., Deutsch (1999) and the study of melodic similarity in Cambouropoulos (2001).

In this article, we are essentially trying to pin a single, quantitative measure to a concept that fundamentally resists such definition. Later, we partly justify this approach with the idea of a consensus truth, but in reality we are forced into the situation out of necessity to build useful applications using current techniques. Before proceeding, however, it is worthwhile to examine in more detail some of the problems that beset the idea of a coherent quantitative measure of music similarity.

Individual Variation

That people have individual tastes and preferences is central to the very idea of music and humanity. By the same token, subjective judgments of the similarity between specific pairs of artists are not consistent between listeners and may vary with an individual's mood or evolve over time. In particular, music that holds no interest for a given subject very frequently "sounds the same."

Multiple Dimensions

The question of the similarity between two artists can be answered from multiple perspectives. Music may be similar or distinct in terms of genre, melody, rhythm, tempo, geographical origin, instrumentation, lyric content, historical timeframe—virtually any property that can be used to describe music. Although these dimensions are not independent, it is clear that different emphases will result in different artists. The fact that both Paul Anka and Alanis Morissette are from Canada might be of paramount significance to a Canadian cultural nationalist, although another person might not find their music at all similar.

Not a Metric

As discussed in Tversky (1977) and elsewhere, subjective similarity often violates the definition of a metric, in particular the properties of symmetry and the triangle inequality. For example, we might say that the 1990s Los Angeles pop musician Jason Falkner is similar to the Beatles, but we would be less likely to say that the Beatles are similar to Jason Falkner, because the more celebrated band serves as a prototype against which to measure.

The triangle inequality can be violated because of the multifaceted nature of similarity: for example, Michael Jackson is similar to the Jackson Five, his Motown roots, and also to Madonna. Both are huge pop stars of the 1980s, but Madonna and the Jackson Five do not otherwise have much in common.

Variability and Span

Few artists are truly a single "point" in any imaginable stylistic space but undergo changes throughout their careers and may consciously span multiple styles within a single album, or even a single song. Trying to define a single distance between any artist and widely ranging, long-lived musicians such as David Bowie or Sting seems unlikely to yield satisfactory results.

Despite all of these difficulties, techniques to automatically determine music similarity have attracted much attention in recent years (Ghias et al. 1995; Foote 1997; Tzanetakis 2002; Logan and Salomon 2001; Aucouturier and Pachet 2002; Ellis et al. 2002). Similarity lies at the core of the classification and ranking algorithms needed to organize and recommend music. Such algorithms could be used in future systems to index vast audio repositories, and thus they must rely on automatic analysis.

Prior Work

Prior work in music similarity has focused on one of three areas: symbolic representations, acoustic properties, and subjective or "cultural" information. We describe each of these below noting in particular their suitability for automatic systems.

Many researchers have studied the music-similarity problem by analyzing symbolic representations such as MIDI music data, musical scores, and the like. A related technique is to use pitch tracking to find a melodic contour for each piece of music. String-matching techniques are then used to compare the transcriptions for each song (e.g., Ghias et al. 1995). However, techniques based on MIDI or scores are limited to music for which this data exists in electronic form, since only limited success has been achieved for pitch tracking of arbitrary polyphonic music.

Acoustic approaches analyze the music content directly and thus can be applied to any music for which one has the audio. Blum et al. (1999) present an indexing system based on matching features such as pitch, loudness, or Mel-Frequency Cepstral Coefficients (MFCCs; these are a compact representation of the frequency spectrum, typically computed over short time windows). Foote (1997) has designed a music indexing system based on histograms of MFCC features derived from a discriminatively trained vector quantizer. Tzanetakis (2002) extracts a variety of features representing the spectrum, rhythm, and chord changes and concatenates them into a single vector to determine similarity. Logan and Salomon (2001) and Aucouturier and Pachet (2002) model songs using local clustering of MFCC features, then determine similarity by comparing the models. Berenzweig, Ellis, and Lawrence (2003) use a suite of pattern classifiers to map MFCCs into an *anchor space*, in which probability models are fit and compared.

With the growth of the World Wide Web, several techniques have emerged that are based on public data derived from subjective human opinion (Cohen and Fan 2000; Ellis et al. 2002). These use text analysis or collaborative filtering techniques to combine data from many users to determine similarity. Because they are based on human opinion, these approaches capture many cultural and other intangible factors that are unlikely to be obtained from audio. The disadvantage of these techniques is that they are only applicable to music for which a reasonable amount of reliable online data is available. For new or undiscovered artists, an audio-based technique may be more suitable.

Acoustic Similarity

In this section, we describe our acoustic-based similarity measures. These are techniques for computing similarity based solely on the audio content, as opposed to subjective measures which involve human judgments. Our techniques fall into a class of methods commonly used in MIR that can be de-

scribed as probabilistic feature modeling and comparison. Essentially, the music is transformed from raw audio samples into a time series of feature vectors, each of which captures the essential characteristics of the sound over a short time interval. The time dimension is then ignored, and the series of feature vectors are considered to be random samples drawn from a probability distribution that represents the piece of music. Probability distributions are much easier to handle when represented with a parameterized class of distributions, and so for each piece of music, the parameters of the chosen probability model are fit to the observed samples that have been extracted from the music. Finally, pieces of music can be compared by comparing the parametric models that have been fit to the audio data.

In fact, this process can operate on the artist level, the album level, or the sub-song level, in addition to the song level. In this article, we use distributions that model an entire artist's work to get an artist-similarity metric, rather than a song similarity metric. The results of each measure are summarized in a *similarity matrix*, a square matrix wherein each entry gives the similarity between a particular pair of artists. The leading diagonal is, by definition, unity, which is the largest value. Sometimes a distance matrix is more convenient, in which entries measure dissimilarity.

In this article, we examine several acoustic-based similarity measures that use the statistical paradigm described above. The techniques are further described in Logan and Salomon (2001) and Berenzweig, Ellis, and Lawrence (2003) and are characterized by the features, models and distance measures used. The next few subsections describe the variants in more detail: first, the feature spaces (MFCC space and anchor space), followed by the modeling techniques and the distance measures (Centroid Distance, Earth Mover's Distance, and the Asymptotic Likelihood Approximation).

Feature Spaces

The first step of audio analysis is to transform the raw audio into a feature space, a numerical representation in which dimensions measure different properties of the input. A good feature space compactly represents the audio, distilling important information and throwing away irrelevant noise. Although many features have been proposed for music analysis, such as spectral centroid, bandwidth, loudness, and sharpness (McKinney and Breebaart 2003), in this article we concentrate on features derived from Mel-Frequency Cepstral Coefficients (MFCCs). These features, originally developed for speech-recognition systems, have been shown to give good performance for a variety of audio classification tasks and are favored by a number of groups working on audio similarity (Blum et al. 1999; Foote 1997; Tzanetakis 2002; Logan 2000; Logan and Salomon 2001; Aucouturier and Pachet 2002; Berenzweig, Ellis, and Lawrence 2003).

The Mel-Cepstrum captures the overall spectral shape, which carries important information about the instrumentation and its timbres, the quality of a singer's voice, and production effects. However, as a purely local feature calculated over a window of tens of milliseconds, it does not capture information about melody, rhythm, or long-term song structure.

We also examine features in an anchor space derived from MFCC features. The anchor space technique is inspired by a folk-wisdom approach to music similarity in which people describe artists by statements such as, "Jeff Buckley sounds like Van Morrison meets Led Zeppelin, but more folksy." Here, musically meaningful categories and well-known anchor artists serve as convenient reference points for describing the music. This idea inspires the anchor space technique, wherein classifiers are trained to recognize musically meaningful categories, and music is subsequently "described" in terms of these categories. Once the classifiers are trained, the audio is presented to each classifier, and the outputs, representing the activation or likelihood of the categories, position the music in the new space.

For this article, we used neural networks as anchor model classifiers, and we used musical genres as the anchor categories, augmented with two supplemental categories. Specifically, we trained a twelve-class network to discriminate between twelve genres: grunge, college rock, country, dance rock, electronica, metal and punk, new wave, rap,

R&B/soul, singer/songwriter, soft rock, and trad rock. Additionally, there were two separate neural nets to recognize the supplemental classes: male/female (sex of the vocalist) and low/high fidelity. Further details about the choice of anchors and the training technique are available in Berenzweig, Ellis, and Lawrence (2003). A system that uses anchor space in a music-browsing application is available online at www.playola.org.

An important point to note is that the input to the classifiers is a large vector consisting of five frames of MFCC vectors plus deltas. This gives the network some time-dependent information from which it can learn about rhythm and tempo, at least on the scale of a few hundred milliseconds.

Modeling and Comparing Distributions

Because feature vectors are computed from short segments of audio, an entire song induces a cloud of points in feature space. The cloud can be thought of as samples from a distribution that characterizes the song, and we can model that distribution using statistical techniques. Extending this idea, we can conceive of a distribution in feature space that characterizes the entire repertoire of each artist.

We use Gaussian mixture models (GMMs) to model these distributions, similar to the technique presented in Logan and Salomon (2001). GMMs are a class of probability models that are often used to model distributions that have more than one mode, or "hump." The classic bell-shaped curve of a single Gaussian is clearly not suited to fitting a distribution with several peaks, and therefore, a weighted mixture of Gaussians is a more powerful modeling tool.

Training mixture models—in other words, fitting the parameters to the observed data—is not as simple as training a single Gaussian, which only entails computing the mean and variance of the data. In fact, iterative procedures must be used to converge to a solution that maximizes the likelihood of the observed data. Several such procedures are commonly used. K-means clustering assigns data points to the nearest cluster, recomputes the cluster centers, and reiterates. The expectation maximization (EM) algorithm is more powerful than K-means, but similar, except data points are given soft (partial) assignments to the clusters.

In this work, two methods of training the Gaussian mixture models were used: simple K-means clustering of the data points to form clusters that were then each fit with a Gaussian component, and standard expectation-maximization (EM) re-estimation of the GMM parameters initialized from the K-means clustering. Although unconventional, the use of K-means to train GMMs without a subsequent stage of EM re-estimation was discovered to be both efficient and useful for song-level similarity measurement in previous work (Logan and Salomon 2001).

The parameters for these models are the mean, covariance, and weight of each cluster. In some experiments, we used a single covariance to describe all the clusters. This is sometimes referred to as a *pooled covariance* in the field of speech recognition; in contrast, an *independent covariance* model estimates separate covariance matrices for each cluster, allowing each to take on an individual shape in feature space, but requiring many more parameters to be estimated from the data.

Having fit models to the data, we calculate similarity by comparing the models. The Kullback-Leibler (KL)-divergence or relative entropy is the natural way to define distance between probability distributions. However, for GMMs, no closed form for the KL-divergence is known. We explore several alternatives and approximations: the centroid distance (Euclidean distance between the overall means); the earth-mover's distance (EMD; see Rubner, Tomasi, and Guibas 1998), which calculates the cost of moving probability mass between mixture components to make them equivalent; and the asymptotic likelihood approximation (ALA) to the KL-divergence between GMMs (Vasconcelos 2001), which segments feature space and assumes only one Gaussian component dominates in each region. Another possibility would be to compute the likelihood of one model given points sampled from the second (Aucouturier and Pachet 2002), but as this is very computationally expensive for large datasets it was not attempted. Computationally, the cen-

troid distance is the cheapest of our methods and the EMD the most expensive.

Subjective Similarity Measures

Whereas acoustic-based techniques can be fully automated, subjective music-similarity measures are derived from sources of human opinion, for instance by mining the World Wide Web. Although these methods cannot always be used on new music because they require observations of human interaction with the music, they can uncover subtle relationships that may be difficult to detect from the audio signal (for example bands that represent the same subculture, are influenced by one another, or even physically look alike).

Subjective measures are also valuable as a "sanity check" against which to evaluate acoustic-based measures; even a sparse ground truth can help validate a more comprehensive acoustic measure. Like the acoustic measures, subjective similarity information can also be represented as a similarity matrix, where the values in each row give the relative similarity between every artist and one target. This section describes several sources of human opinion about music similarity and how to convert them into a useful similarity measure.

Survey

The most straightforward way to gather human similarity judgments is to explicitly ask for them in a survey. We have previously constructed a Web site, musicseer.com, to conduct such a survey (Ellis et al. 2002). We defined a set of some 400 popular artists (described in a subsequent section), then presented subjects with a list of 10 artists ($a_1, \ldots, a_{10}$), and a single target artist a_t, and asked "Which of these artists is most similar to the target artist?" We interpret each response to mean that the chosen artist a_c is more similar to the target artist a_t than any of the other artists in the list only if those artists are known to the subject, which we can infer by seeing if the subject has ever selected the artists in any context.

Ideally, the survey would provide enough data to derive a full similarity matrix, for example by counting how many times users selected artist a_i being most similar to artist a_j. However, even with the 22,000 responses collected, the coverage of our modest artist set is relatively sparse: only around 7.5% of all our artist pairs were directly compared, and only 1.7% of artist pairs were ever chosen as most similar. We constructed this sparse similarity matrix by populating each row with the number of times a given artist was chosen as most similar to a target as a proportion of the trials in which it could have been chosen. This heuristic worked quite well for our data.

Expert Opinion

Rather than surveying the masses, we can ask a few experts. Several music-related online services contain music taxonomies and articles containing similarity data. The All Music Guide (www.allmusic.com) is such a service in which professional editors write brief descriptions of a large number of popular musical artists, often including a list of similar artists. We extracted the "similar artists" lists from the All Music Guide for the 400 artists in our set, discarding any artists from outside the set, resulting in an average of 5.4 similar artists per list (so 1.35% of artist pairs had direct links). Twenty-six of our artists had no neighbors from within the set.

As in Ellis et al. (2002), we convert these descriptions of the immediate neighborhood of each artist into a similarity matrix by computing the path length between each artist in the graph where nodes are artists and there is an edge between two artists if the All Music Guide editors consider them similar. Our construction is symmetric, because links between artists were treated as nondirectional. We call this the Erdös measure, after the technique used among mathematicians to gauge their relationship to Paul Erdös. This extends the similarity measure to cover 87.4% of artist pairs.

Playlist Co-Occurrence

Another source of human opinion about music similarity is human-authored playlists, such as those selected for a mixed tape or compilation CD. Our assumption is that songs co-occurring in the same playlist will, on average, be more similar than two randomly chosen songs. This assumption is suspect for many types of playlists, but as we will see it proves useful. The Web is a rich source for such playlists. In particular, we gathered around 29,000 playlists from The Art of the Mix (www.artofthemix.org), a Web site that serves as a repository and community center for playlist hobbyists.

To convert this data into a similarity matrix, we begin with the normalized playlist co-occurrence matrix, where entry (i,j) represents the joint probability that artist a_i and a_j occur in the same playlist. However, this probability is influenced by overall artist popularity, which should not affect a similarity measure. Therefore, we use a normalized conditional probability matrix instead: entry (i,j) of the normalized conditional probability matrix C is the conditional probability $p(a_i \mid a_j)$ divided by the prior probability $p(a_i)$. Because

$$c_{ij} = \frac{p(a_i \mid a_j)}{p(a_i)} = \frac{p(a_i, a_j)}{p(a_i)p(a_i)}$$

this is an appropriate normalization of the joint probability. Note that the expected logarithm of this measure is the mutual information $I(a_i;a_j)$ between artist a_i and a_j.

Using the playlists gathered from Art of the Mix, we constructed a similarity matrix with 51.4% coverage for our artist set (i.e., more than half of the matrix cells were nonzero).

User Collections

Similar to user-authored playlists, individual music collections are another source of music similarity often available on the Internet. Mirroring the ideas underlying collaborative filtering, we assume that artists co-occurring in someone's collection have a better-than-average chance of being similar, which increases with the number of co-occurrences observed.

We retrieved user collection data from OpenNap, a popular music-sharing service, although we were careful not download any audio files. After discarding artists not in our data set, we were left with about 176,000 user-to-artist relations from about 3,200 user collections. To turn this data into a similarity matrix, we used the same normalized conditional probability technique for playlists as described above. This returned a similarity matrix with nonzero values for 95.6% of the artist pairs.

"Webtext"

A rich source of information resides in text documents that describe or discuss music. Using techniques from the IR community, we derived artist-similarity measures from documents returned from Web searches (Whitman and Lawrence 2002). The best-performing similarity matrix from that study, which measures document similarity based on frequency bigram phrases, is used here. This matrix has essentially full coverage.

Evaluation Methods

In this section, we describe our evaluation methodology. First, a caveat: any evaluation system inherently assumes some idea of ground truth against which the candidate is evaluated. Although similarity is inherently subjective, thus there is no authoritative ground truth, we can tentatively treat the subjective data described above as if it were ground truth. This approach is partly justified because the data are derived from human choices, but more importantly, we later leverage the diversity of sources to examine how well the sources agree with each other.

We present several techniques for evaluating a similarity measure. The first technique is a general method for evaluating one similarity matrix given another as a reference ground truth. Then we present two techniques specifically designed for using the survey data as ground truth.

Evaluation Against a Reference Similarity Matrix

If we are given one similarity metric as ground truth, how can we calculate the agreement achieved by other similarity matrices? We use an approach inspired by practice in text information retrieval (Breese, Heckerman, and Kadie 1998). Each matrix row is sorted into decreasing similarity and treated as the results of a query for the corresponding target artist. The top N "hits" from the reference matrix define the ground truth, with exponentially decaying weights so that the top hit has weight 1, the second hit has weight α_r, the next $(\alpha_r)^2$ etc. (We consider only N hits to minimize issues arising from similarity information sparsity.) The candidate matrix "query" is scored by summing the weights of the hits by another exponentially decaying factor, so that a ground-truth hit placed at rank r is scaled by $(\alpha_c)^r$. Formally, we define the *Top-N Ranking Agreement Score* for row i as:

$$s_i = \sum_{r=1}^{N} (\alpha_r)^r (\alpha_c)^{k_r}$$

where k_r is the ranking according to the candidate measure of the r^{th}-ranked hit under the ground truth. The parameters α_c and α_r govern how sensitive the metric is to ordering under the candidate and reference measures, respectively. We used the values $\alpha_r = 0.5^{1/3}$ and $\alpha_c = (\alpha_r)^2$ to emphasize the position of the top few ground-truth hits. With $N = 10$ and these values of α_r and α_c, the optimal score—achieved when the top ten ground truth hits are the same, and in the same order as, the top ten from the candidate matrix—is 0.999. Finally, the overall score for the experimental similarity measure is the average of the normalized row scores

$$S = \frac{1}{N} \sum_{i}^{N} s_i / s_{\max}$$

where $S_{\max}$ is the optimal score. Thus a larger rank agreement score is better, with 1.0 indicating perfect agreement.

One issue with this measure arises from the handling of ties. Because much of the subjective information is based on counts, ranking ties are not uncommon (an extreme case being the 26 "disconnected" artists in the expert measure, who must be treated as uniformly dissimilar to all artists). We handle this by calculating an average score over multiple random permutations of the equivalently-ranked entities; owing to the interaction with the top-N selection, a closed-form solution has eluded us. The number of repetitions was based on empirical observations of the variation in successive estimates to obtain a stable estimate of the underlying mean.

Evaluation Against Survey Data

The similarity data collected using our Web-based survey can be argued to be a good independent measure of ground-truth artist similarity, because users were explicitly asked to indicate similarity. However, the coverage of the similarity matrix derived from the survey data is only about 1.7%, which makes it undesirable to use as a ground-truth reference as described in the previous section. Instead, we can compare the individual user judgments from the survey directly to the metric we wish to evaluate. That is, we ask the similarity metric the same questions that we asked the users and compute an average agreement score.

We used two variants of this idea. The first, *average response rank*, determines the average rank of the artists chosen from the list of ten presented in the survey according to the experimental metric. For example, if the experimental metric agrees perfectly with the human subject, then the ranking of the chosen artist will be first in every case, whereas a random ordering of the artists would produce an average ranking of 5.5. In practice, the ideal score of 1.0 is not possible, because survey subjects did not always agree on artist similarity; therefore, a ceiling exists corresponding to the single, consistent metric that optimally matches the survey data. For our data, this was estimated to give a score of 2.13.

The second approach is simply to count how many times the similarity measure agrees with the user about the first-place (most similar) artist from the list. This proportion, called *first-place agreement*, has the advantage that it can be viewed as

the average of a set of independent binomial (binary-valued) trials, meaning that we can use a standard statistical significance test to confirm that certain variations in values for this measure arise from genuine differences in performance, rather than random variations in the measure. Our estimate of the best possible first-place agreement with the survey data was 53.5%.

Multi-Site Evaluation Procedures

To compare results between research sites, it is necessary to have a common database and a common evaluation method. Using the evaluation techniques described above, we had to share data between centers. However, we encounter legal restrictions when attempting to share copyrighted music. Although efforts are underway to procure a database of music that is free for use in the music information retrieval community (Downie 2002), negotiations can be slow, and for certain types of research it is not necessary to have the full audio. For our purposes, it suffices to share the MFCC features derived from the audio, since all of the acoustic-based similarity measures we use begin with computing the MFCCs. Similarly, other researchers wanting to experiment with different techniques need only share the front-end features.

Because our audio experiments were conducted at two sites, a level of discipline was required when setting up the data. We shared MFCC features rather than raw audio, both to save bandwidth and to avoid copyright problems, as mentioned. This had the added advantage of ensuring both sites started with the same features when conducting experiments. Duplicated tests on a small subset of the data were used to verify the equivalence of our processing and scoring schemes.

We believe that this technique of establishing common feature-calculation tools, then sharing common feature sets, could be useful for future cross-group collaborations and should be seriously considered by those proposing evaluations, and we would be interested in sharing our derived features.

We have compiled a relatively large dataset from audio and online sources. The dataset covers 400 artists chosen to have the maximal overlap of the user collection (OpenNap) and playlist (The Art of the Mix) data. We had previously purchased audio corresponding to the most popular OpenNap artists and had also used these artists to construct the survey data. For each artist, our database contains audio, survey responses, expert opinions from All Music Guide, playlist information, OpenNap collection data, and Webtext data.

The audio data consists of 707 albums and 8,772 songs, for an average of 22 songs per artist. The specific track listings for this database, which we refer to as "uspop2002," are available online at www.ee.columbia.edu/~dpwe/research/musicsim.

Experiments and Results

A number of experiments were conducted to answer the following questions about acoustic- and subjective-based similarity measures. First, is anchor space better for measuring similarity than MFCC space? Second, which method of modeling and comparing feature distributions is best? Third, which subjective similarity measure provides the best ground truth, e.g., in terms of agreeing best, on average, with the other measures?

Although it risks circularity to define the best ground truth as the measure that agrees best with the others, we argue that because the various measures are constructed from diverse data sources and methods, any correlation between them should reflect a true underlying consensus among the people who generated the data. A measure consistent with all these sources must approach a "consensus truth," even if no absolute ground truth actually exists.

Acoustic Similarity Measures

We first compare the acoustic-based similarity measures, examining artist models trained on MFCC and anchor space features. Each model is trained using features calculated from the available audio for that artist. Our MFCC features are 20-dimensional and are computed using 32-msec

frames overlapped by 16 msec. The anchor space features have 14 dimensions where each dimension represents the posterior probability of a pre-learned acoustic class given the observed audio as described in the section "Acoustic Similarity" above.

In a preliminary experiment, we performed dimensionality reduction on the MFCC space by taking the first 14 dimensions of a PCA analysis and compared results with the original 20-dimensional MFCC space. There was no appreciable difference in results, confirming that any difference between the anchor-based and MFCC-based models is not owing to the difference in dimensionality.

Table 1 shows results for similarity measures based on MFCC space, in which we compare the effect of varying the distribution models and the distribution similarity method. For the GMM distribution models, we vary the number of mixtures, use pooled or independent variance models, and train using either plain K-means, or K-means followed by EM re-estimation. Distributions are compared using centroid distance, ALA, or EMD (as described in the section "Modeling and Comparing Distributions"). We also compare the effect of including or excluding the first cepstral coefficient, c_0, which measures the overall intensity of a signal. Table 1 shows the average response rank and first-place agreement percentage for each approach.

From this table, we see that the different training techniques for GMMs give comparable performance and that more mixture components help up to a point. Pooling the data to train the covariance matrices is useful, as has been shown in speech recognition, because it allows for more robust covariance parameter estimates. Omitting the first cepstral coefficient gives better results, possibly because similarity is more related to spectral shape than overall signal energy, although this improvement is less pronounced when pooled covariances are used. The best system is one that uses pooled covariances and ignores c_0. Models trained with the simpler K-means procedure appear to perform as well as GMMs and thus are preferred.

A similar table was constructed for anchor-space-based methods, which revealed that full, independent covariance using all 14 dimensions was the best-performing method. Curiously, while the ALA distance measure performed poorly on MFCC-based models, it performed competitively with EMD on anchor-space models. We are still investigating the cause; perhaps it is because the assumptions behind the asymptotic likelihood approximation do not hold in MFCC space.

Table 1. Average response rank and first-place agreement percentages for various similarity schemes based on MFCC features. Lower values are better for average response rank, and larger percentages are better for first-place agreement

			Independent		*Pooled*		
	#mix	*c0?*	*ALA*	*EMD*	*ALA*	*Cntrd*	*EMD*
EM	8	y	4.76 / 16%	4.46 / 20%	4.72 / 17%	4.66 / 20%	4.30 / 21%
	8	n	–	4.37 / 22%	–	–	4.23 / 22%
	16	n	–	4.37 / 22%	–	–	4.21 / 21%
K-means	8	y	–	4.64 / 18%	–	–	4.30 / 22%
	8	n	4.70 / 16%	4.30 / 22%	4.76 / 17%	4.37 / 20%	4.28 / 21%
	16	y	–	4.75 / 18%	–	–	4.25 / 22%
	16	n	4.58 / 18%	4.25 / 22%	4.75 / 17%	4.37 / 20%	4.20 / 22%
	32	n	–	–	4.73 / 17%	4.37 / 20%	4.15 / 23%
	64	n	–	–	4.73 / 17%	4.37 / 20%	4.14 / 23%
Optimal			2.13 / 53.5%				
Random			5.50 / 11.4%				

The comparison of the best-performing MFCC and anchor-space models is shown in Table 2. We see that both have similar performance under these metrics, despite the prior information encoded in the anchors.

Table 2. Best-in-class comparison of anchor versus MFCC-based measures (average response rank / first-place agreement percentage)

	MFCC	*Anchor*
#mix	*EMD*	*ALA*
8	4.28 / 21.3%	4.25 / 20.2%
16	4.20 / 22.2%	4.20 / 19.8%

The MFCC system uses K-means training, pooled diagonal covariance matrices, and excludes c_0. The anchor-space system uses EM training, independent full covariance matrices, and includes c_0.

Comparing Ground Truth Measures

Now we turn to a comparison of the acoustic and subjective measures. We take the best-performing approaches in each feature-space class (MFCC and anchor space, limiting both to 16 GMM components for parity) and evaluate them against each of the subjective measures. At the same time, we evaluate each of the subjective measures against each other. The results are presented in Table 3. Rows represent similarity measures being evaluated, and the columns give results treating each of our five subjective similarity metrics as ground truth. Top-*N* ranking agreement scores are computed as described in the section ''Evaluation Against a Reference Similarity Matrix.''

The means down each column, excluding the self-reference diagonal, are also shown (denoted ''mean⋆''). The column means can be taken as a measure of how well each measure approaches ground truth by agreeing with all the data. By this standard, the survey-derived similarity matrix is best, but its very sparse coverage makes it less useful. The user collection (''opennap'') data has the second-highest mean⋆, including particularly high agreement with the survey metric, as can be seen when the Top-*N* ranking agreements are plotted as an image in Figure 1. Thus, we consider the user collections as the best single source of a ground truth based on this evidence, with the survey's (and hence the first-place agreement metric's) providing reliable data also. (Interestingly, the collection data does less well agreeing with the survey data when measured by the first-place agreement percentage; we infer that it is doing better at matching further down the rankings.)

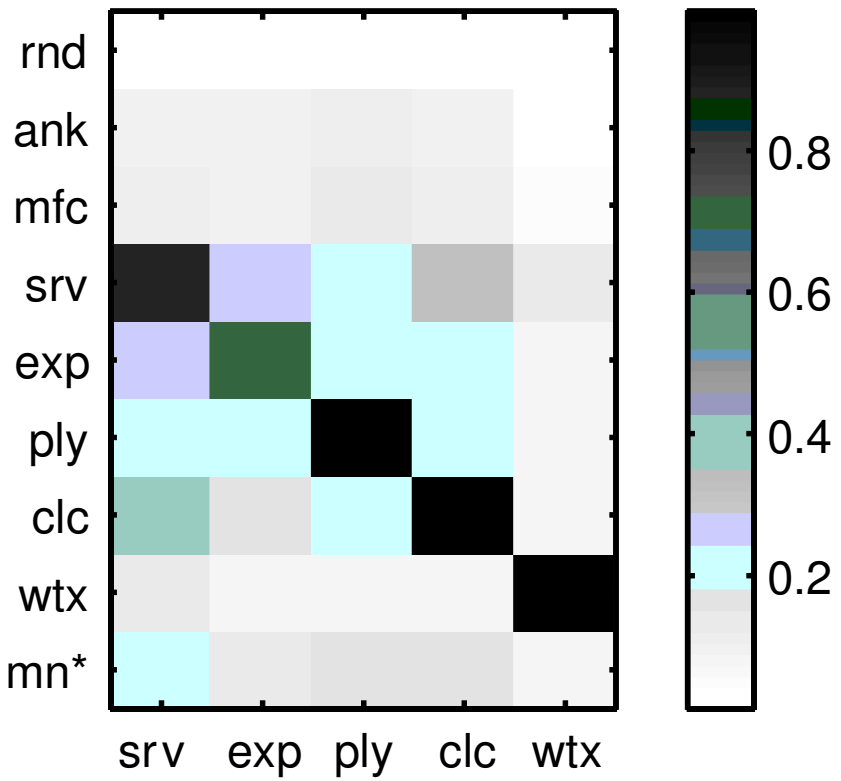
The natural asymmetry of Table 3 exists because $\alpha_c \neq \alpha_r$, and the diagonal is less than one because of the randomized tiebreakers necessary owing to the

Table 3. First-place agreement percentages (with survey data) and Top-*N* ranking agreement scores (against each candidate's ground truth) for acoustic and subjective similarity measures

	1stplace	*Survey*	*Expert*	*Playlist*	*Collection*	*Webtext*
Random	11.8%	0.015	0.020	0.015	0.017	0.012
Anchor	19.8%	0.092	0.095	0.117	0.097	0.041
MFCC	22.2%	0.112	0.099	0.142	0.116	0.046
Survey	53.5%	**0.874**	0.249	0.204	0.331	0.121
Expert	27.9%	0.267	**0.710**	0.193	0.182	0.077
Playlist	26.5%	0.222	0.186	**0.985**	0.226	0.075
Collection	23.2%	0.355	0.179	0.224	**0.993**	0.083
Webtext	18.5%	0.131	0.082	0.077	0.087	**0.997**
mean⋆		0.197	0.148	0.160	0.173	0.074

The mean⋆ rows represent the means of each ground-truth column, excluding the bolded ''cheating'' diagonal and the ''random'' row.

*Figure 1. Top-*N *ranking agreement scores from Table 3 plotted as a grayscale image.*



sparsity of the sources. If we calculate a symmetric agreement score for each pair of sources by averaging the two asymmetric numbers, some interesting results emerge. The best-agreeing pair is the survey and the collection data, which is somewhat surprising given the very different nature of data sources: explicit user judgments in the survey and co-occurrence of artists in user collections. Less surprising is the agreement between the survey and expert sources, which both come from explicit judgments by humans, and between the collection and the playlist sources, which both are derived from co-occurrence data.

We mentioned that a key advantage of the first-place agreement measure was that it allowed the use of established statistical significance tests. Using a one-tailed test under a binomial assumption, first-place agreements differing by more than about 1% are significant at the 5% level for this data (10,884 trials). Thus, all the subjective measures show significantly different results, although differences among the variants in modeling schemes from Tables 1 and 2 are at the edge of significance.

Conclusions and Future Work

Returning to the three questions posed in the previous section, based on the results shown above, we draw several conclusions. First, MFCC and anchor space achieve comparable results on the survey data. Second, K-means training is comparable to EM training. Using pooled, diagonal covariance matrices is beneficial for MFCC space, but in general the best modeling scheme and comparison method depend on the feature space being modeled. Third, the measure derived from co-occurrence in personal music collections is the most useful ground

truth, although some way of combining the information from different source warrants investigation since they are providing different information.

The work covered by this article suggests many directions for future research. Although the acoustic measures achieved respectable performance, there is still much room for improvement. One glaring weakness of our current features is their failure to capture any temporal structure information, although it is interesting to see what can be achieved based on this limited representation.

Based on our cross-site experience, we feel that this work points the way to practical music-similarity system evaluations that can even be carried out on the same database, and that the serious obstacles to sharing or distributing large music collections can be avoided by transferring only derived features (which should also reduce bandwidth requirements). To this end, we have set up a web site giving full details of our ground truth and evaluation data (www.ee.columbia.edu/~dpwe/research/musicsim). We will also share the MFCC features for the 8,772 tracks we used in this work by burning DVDs to send to interested researchers. We are also interested in proposals for other features that it would be valuable to calculate for this data set.

Acknowledgments

We are grateful for support for this work received from NEC Laboratories America, Inc. We also thank the anonymous reviewers for their useful comments.

Much of the content of this article also appears in our white paper presented at the Workshop on the Evaluation of Music Information Retrieval (MIR) Systems at SIGIR-03, Ottawa, August 2003.

References

Aucouturier, J. J., and F. Pachet. 2002. "Music-Similarity Measures: What's the Use?" *Proceedings of the Third International Symposium on Music Information Retrieval*. Paris: IRCAM, pp. 157–163.

Berenzweig, A., D. P. W. Ellis, and S. Lawrence. 2003. "Anchor Space for Classification and Similarity Measurement of Music." In *Proceedings of the 2003 IEEE International Conference on Multimedia and Expo*. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Blum, T. L., et al. 1999. "Method and Article of Manufacture for Content-Based Analysis, Storage, Retrieval, and Segmentation of Audio Information." U.S. Patent No. 5,918,223.

Breese, J. S., D. Heckerman, and C. Kadie. 1998. "Empirical Analysis of Predictive Algorithms for Collaborative Filtering." *Fourteenth Annual Conference on Uncertainty in Artificial Intelligence*. San Francisco, California: Morgan Kaufmann, pp. 43–52.

Cambouropoulos, E. 2001. "Melodic Cue Abstraction, Similarity, and Category Formation: A Computational Approach." *Music Perception* 18(3):347–370.

Cohen, W.-W., and W. Fan. 2000. "Web-Collaborative Filtering: Recommending Music by Crawling the Web." *WWW9 / Computer Networks* 33(1–6):685–698.

Deutsch, D., ed. 1999. *The Psychology of Music*, 2nd ed. New York: Academic Press.

Downie, J. S., ed. 2002. *The MIR/MDL Evaluation Project White Paper Collection*, 2nd ed. Champaign, Illinois: GSLIS.

Ellis, D. P., et al. 2002. "The Quest for Ground Truth in Musical Artist Similarity." *Proceedings of the Third International Symposium on Music Information Retrieval*. Paris: IRCAM, pp. 170–177.

Foote, J. 1997. "Content-Based Retrieval of Music and Audio." In C. J. Kuo, S. Chang, and V. N. Gudivada, eds. *SPIE Vol. 3229: Multimedia Storage and Archiving Systems II*. Bellingham, Washington: SPIE Press, pp. 138–147.

Ghias, A., et al. 1995. "Query by Humming." *ACM Multimedia* 95:231–236.

Goldstone, R. L., D. Medin, and D. Gentner. 1991. "Relational Similarity and the Nonindependence of Features in Similarity Judgments." *Cognitive Psychology* 23:222–262.

Logan, B. 2000. "Mel-Frequency Cepstral Coefficients for Music Modeling." *Proceedings of the First International Symposium on Music Information Retrieval* 2000. Amherst: University of Massachusetts at Amherst, n.p.

Logan, B., and A. Salomon. 2001. "A Music-Similarity Function Based on Signal Analysis." Paper presented at the 2001 International Conference on Multimedia and Expo, Tokyo, Japan, 25 August.

McKinney, M. F., and J. Breebaart. 2003. "Features for Audio and Music Classification." *Proceedings of the Third International Symposium on Music Information Retrieval*. Paris: IRCAM, pp. 151–158.
Rubner, Y., C. Tomasi, and L. Guibas. 1998. "A Metric for Distributions with Applications to Image Databases." *Proceedings of the Sixth International Conference on Computer Vision*. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, p. 59.
Tversky, A. 1977. "Features of Similarity." *Psychological Review* 84(4):327–352.
Tzanetakis, G. 2002. "Manipulation, Analysis, and Retrieval Systems for Audio Signals." Ph.D. Thesis, Princeton University.
Vasconcelos, N. 2001. "On the Complexity of Probabilistic Image Retrieval." *Proceedings of the Eighth International Conference on Computer Vision*, vol. 2. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, pp. 400–407.
Whitman, B., and S. Lawrence. 2002. "Inferring Descriptions and Similarity for Music from Community Metadata." *Proceedings of the 2002 International Computer Music Conference*. San Francisco, California: International Computer Music Association, pp. 591–598.